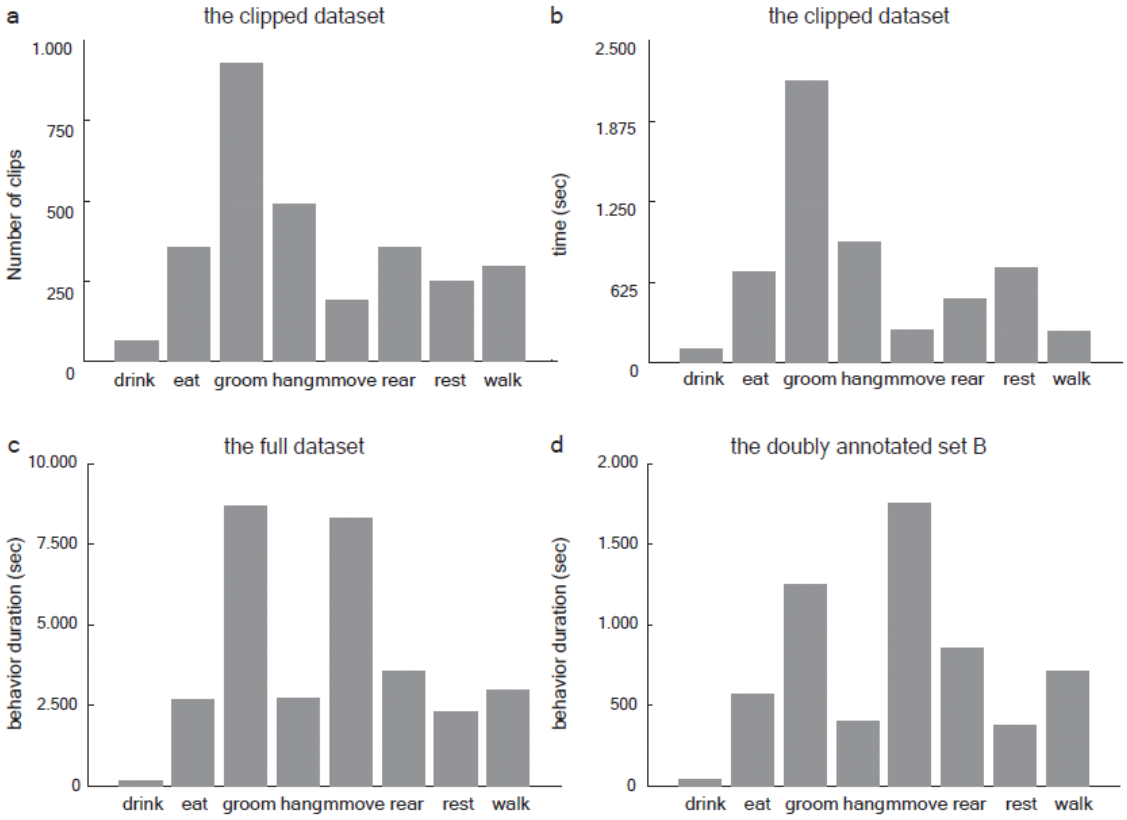
AUTOMATED HOME-CAGE BEHAVIORAL PHENOTYPING OF MICE

Hueihan Jhuang[1], Estibaliz Garrote[1], Xinlin L. Yu[3], Vinita Khilnani[3], Tomaso Poggio[1] and Andrew D. Steele[3] and Thomas Serre[1, 2]

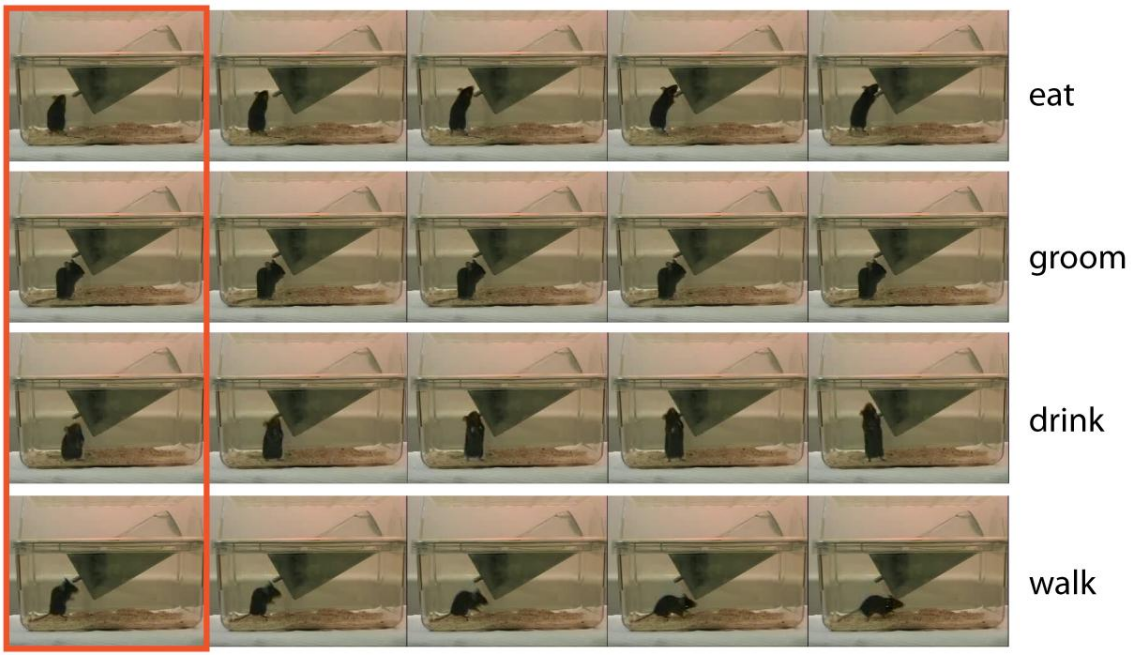
The Supplementary information contains:

- Supplementary Figures (S1, S2, S3)
- Supplementary Tables (S1, S2)
- Supplementary Note
- Supplementary Methods
- Supplementary References

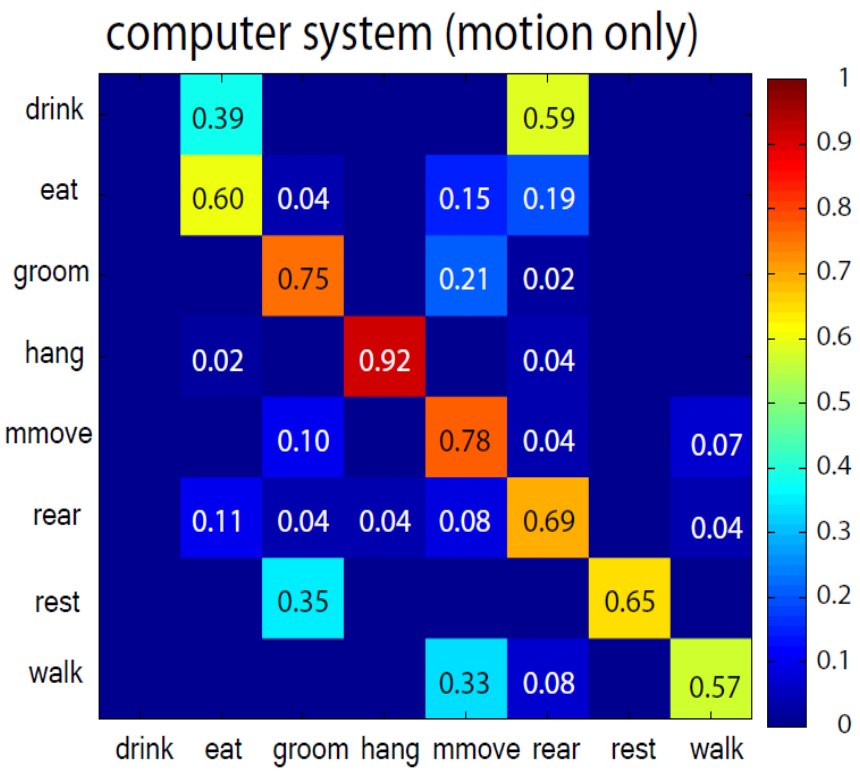
SUPPLEMENTARY FIGURES



Supplementary Figure S1: Distribution of behavior annotations for the '*clipped database'* over **(a)** the number of clips and **(b)** total time. Distribution of behavior annotations for **(c)** the '*full database'* annotated by '*Annotator group 1*' and for **(d)** the '*set B'*, which corresponds to a subset of the full database that was annotated by both '*Annotator group 1*' and '*Annotator group 2'* to evaluate the agreement between two sets of independent annotators.



Supplementary Figure S2: Single frames are ambiguous. Each row corresponds to a short video clip. While the leftmost frames (red bounding box) all look quite similar, they each correspond to a different behavior (text on the right side). Because of this ambiguity, frame-based behavior recognition is unreliable and temporal models that integrate the local temporal context over adjacent frames are needed for robust behavior recognition.



Supplementary Figure S3: The confusion matrix evaluated on the doubly annotated '*set B*' to measure the agreement of a system using motion features with human manual scoring. To reduce clutter and improve the clarity of the presentation, entries with values smaller than 0.01 are not shown. The color bar indicates the percent agreement, with more intense shades of red indicating strong agreement and lighter shades of blue indicating weaker agreement

SUPPLEMENTARY TABLES

	Symbol/ Formula	Meaning
1	Cx	x coordinate of the mouse center
2	Cy	y coordinate of the mouse center
3	w	width of the mouse
4	h	height of the mouse
5	h/w	aspect ratio of the mouse
6	fd	closet distance between the mouse and the feeder
7	$td1$	closet distance between the mouse mouth and the tip of the drinking tube
8	$td2$	closet distance between the mouse mouth and the point of the drinking tube that attaches the food hopper
9	$Vx(t) = \lvert Cx(t) - Cx(t-1) \rvert$	speed of the mouse in the x direction
10	$Vy(t) = \lvert Cy(t) - Cy(t-1) \rvert$	speed of the mouse in the y direction

Supplementary Table S1: A list of 10 position- and velocity-based features. Feature # 7 is close to zero for the 'drinking' behavior; this helps distinguish it from the rest of behaviors. However, when the mouse eats from near the drinking tube, Feature # 7 is also close to zero for the 'eating' behavior. In order to disambiguate these two behaviors, we introduce Feature # 8: for eating, it is close to zero, while for drinking, it equals to the length of the tube. See Fig S2, the first row ('eating') and the third row ('drinking') of the last column for an example.

HCS label	System label
Drink	Drink
Chew	Eat
Eat	
Groom	Groom
Hang Cuddled	Hang
Hang Vertically	
Hang Vertically From Hang Cuddled Hang	
Hang Vertically From Rear Up	
Remain Hang Cuddled	
Remain Hang Vertically	
Awaken	Micro-move
Pause	
Remain Low	
Sniff	
Twitch	
Come Down	Rear
Come Down From Partially Reared	
Come Down To Partially Reared	
Stretch Body	
Land Vertically	
Rear Up	
Rear up From Partially Reared	
Rear up To Partially Reared	
Remain Partially Reared	
Remain Rear Up	
Sleep	Rest
Stationary	
Circle	Walk
Turn	
Walk Left	
Walk Right	
Walk Slowly	
Dig	Unknown Behavior
Forage	
Jump	
Repetitive Jumping	
Unknown Behavior	
Urinate	

Supplementary Table S2: The 1st column of the table shows a list of 38 labels used by HomeCageScan system evaluation. Our system chose the eight types of action categories (2rd column) as a proof-of-concept because these eight types represent almost the entirety of actions in the home cage. We lumped the behaviors used in Clever Sys. to compare the two systems fairly on the same eight types of actions.

SUPPLEMENTARY NOTE

Sensor-based approaches. Previous automated systems (e.g., ref. [8, 9, 11, 12, 32]) have relied for the most part on the use of sensors to monitor behavior. Popular sensor-based approaches include the use of PVDF sensors [33], infrared sensors [9, 34-36], RFID transponders [37] as well as photobeams [8]. Such approaches have been successfully applied to the analysis of coarse locomotion activity as a proxy to measure global behavioral states such as active vs. resting. Other studies have successfully used sensors for the study of food and water intake [32, 38]. However the physical measurements obtained from these sensor-based approaches limit the complexity of the behavior that can be measured. This problem remains even for commercial systems using transponder technologies such as the IntelliCage system (NewBehavior Inc). While such systems can be effectively used to monitor the locomotion activity of an animal as well as other pre-programmed activities via operant conditioning units located in the corners of the cage, such systems alone cannot be used to study natural behaviors such as grooming, sniffing, rearing or hanging, etc.

Video-based approaches. One of the possible solutions to address the problems described above is to rely on vision-based techniques. In fact such approaches are already bearing fruit for the automated tracking [18, 19, 39] and recognition of behaviors in insects [20, 21]. Several open-source and commercial computer-vision systems for the tracking of rodents have been developed [12, 40-45]. As for sensor-based approaches, such systems are particularly suitable for studies involving coarse locomotion activity based on spatial measurements such as the distance covered by an animal or its speed [46-49]. Video-tracking based approaches tend to be more flexible and much more cost efficient. However, as in the case of sensor-based approaches, these systems alone are not suitable for the analysis of fine animal activities such as grooming, sniffing, rearing or hanging.

The first effort to build an automated computer vision system for the monitoring of rodent behavior was initiated at USC. As part of this SmartVivarium project [50], an initial computer-vision system was developed for both the tracking [40] of the animal as well as the recognition of five behaviors (eating, drinking, grooming, exploring and resting, see ref. [23]). Xue & Henderson recently described an approach [22, 51] for the analysis of rodent

behavior however the system was only tested on synthetic data [52] and a very limited number of behaviors. Overall, none of the existing systems [22, 23, 51] have been tested in a real-world lab setting using long uninterrupted video sequences containing potentially ambiguous behaviors or at least evaluated against human manual annotations on large databases of video sequences using different animals and different recording sessions. Recently a commercial system (HomeCageScan by CleverSys, Inc) was also introduced and the system was successfully used in several behavioral studies [5-8]. Such commercial products typically rely on the animal contour and relatively simple heuristics such as the position of the animal in the cage to infer behavior. They thus remain limited in their scope (tracking of simple behaviors) and error-prone (see ref. [6] and Table 1 for a comparison against our manual annotations). In addition, the software packages are proprietary: there is no simple way for the end user to improve its performance or to customize it to specific needs.

SUPPLEMENTARY METHODS

System Overview. Figure 2 provides an overview of the computer vision system used. The system takes as input a video sequence recorded from a video camera. It then converts every frame of a video sequence into a representation, which is suitable for the recognition of complex behaviors. This representation is a feature vector that consists of motion-based as well as position- and velocity-based features. For the motion features, each coefficient of the vector corresponds to the degree of similarity between low-level motion at the current frame and stored space-time motion templates learned from the set of behaviors of interest (*'clipped database'*, see main text). An action label is then obtained for every frame of a video by passing this feature vector to a temporal model for classification. The temporal model used here is a hidden Markov Support Vector Machine (SVMHMM) [28, 29, 53, 54], which is an extension of the popular Support Vector Machine classifier developed by Vapnik [55] in the 90's, for sequence tagging. This temporal model was trained using manually annotated examples extracted from the video database denoted *'full database'* in the main text. This database involved labeling every frame for 12 distinct videos (from different mice recorded in different conditions) for a

total of 10.4 hours of annotated video. The output of the system is thus a label corresponding to a specific behavior of interest for every frame of a video sequence.

The learning of the basic dictionary of space-time motion templates as well as the feature computation and the learning of the temporal model are described in detail in the following sections.

Computing motion features The approach taken here builds directly on the work by Jhuang et al. [25] (which itself builds on the work by Giese & Poggio [24]). The system is based on the organization of the dorsal stream of the visual cortex and was shown to compete with state-of-the-art computer vision systems. The system is organized hierarchically: low-level features are first extracted at the bottom layer and progressively transformed, through successive ***S*** and ***C*** stages of processing, to become increasingly complex and invariant (see ref [25] for details). In the ***S***1 stage, feature maps are obtained by convolving an input sequence with four spatio-temporal filters (each is 9 pixels × 9 pixels × 9 frames) tuned to four different directions of motion. The four directions are equally spaced between 0 and $\pi/2$. This linear stage is followed by a non-linear contrast normalization whereby the filter response is divided by the L1 norm of the corresponding (9 pixels × 9 pixels × 9 frames) video patch. This results in four ***S***1 maps for every input frame where each map corresponds to one direction and each value on the map corresponds to the amount of motion present in this direction. The nature of the non-linearity contrast normalization (i.e., L1 vs. L2 vs. no-normalization), the number of motion directions used and the resolution of the input video sequences were carefully optimized in a preliminary experiment carried on a subset of the clipped database.

Beyond this initial ***S***1 stage, processing is then hierarchical: alternating between a template matching (***S*** layers) and a max pooling operation (***C*** layers) gradually increases feature complexity and translation invariance. At the ***C***1 stage some tolerance to small spatial deformations is obtained via a local max operation over neighborhing pixels (8 × 8 pixels) of each ***S***1 map. Note the local-max is performed across ***x- y*** space, not directions, therefore we obtain four ***C***1 maps for each input frame. Next, for the four ***C***1 maps, a template matching against ***d*** space-time motion templates (or called alternatively as 'motion prototype') is performed, creating ***d S***2 maps for every input frame. More precisely, at each spatial location of the ***C***1 maps, a patch centering at the location and

containing the four directions is compared to each of the $\boldsymbol{d}$ prototype patches, the resulting similarity is stored as a $\boldsymbol{S}2$ value. At the top of the hierarchy, a vector of $\boldsymbol{d}$ position invariant $\boldsymbol{C}2$ features for each input frame is obtained by computing a global max for each of the $\boldsymbol{d}$ $\boldsymbol{S}2$ maps.

Learning and selecting the space-time motion templates Each motion prototype corresponds to a 3-dimensional patch of size $\boldsymbol{4} \times \boldsymbol{n} \times \boldsymbol{n}$ (direction × pixel × pixel) obtained by cropping a $\boldsymbol{n} \times \boldsymbol{n}$ ($\boldsymbol{n}$ = 4, 8, 12, 16) patch of four $\boldsymbol{C}1$ maps computed at a training frame. These $\boldsymbol{d}$ prototype represent the intermediate-level features of the model, and are sampled from random locations of $\boldsymbol{C}1$ maps of randomly drawn training images (*'clipped database'*) in an initial feature-learning stage. To select a more discriminative dictionary of motion prototypes (and speed up the overall system), we applied a feature selection technique called zero-norm SVM [56] on the initial set of $\boldsymbol{d'}$ ($\boldsymbol{d'} >> \boldsymbol{d}$) $\boldsymbol{C}2$ features. We firstly select a random subset of frames from the training set (*'clipped database'*), and compute the $\boldsymbol{d'}$-dimensional $\boldsymbol{C}2$ features for each frame by a matching against an initial set of $\boldsymbol{d'}$ prototypes. Selection was then done in an iterative manner: in each round, a SVM classifier was trained on the pool of $\boldsymbol{C}2$ features; for every training point, each of the $\boldsymbol{d'}$ dimensions was re-weighted using the weights of the hyper-plane returned by the SVM. Typically this leads to sparser hyper-plane weights at each round, and eventually leading to a final set of 300 dimensions with strong weights, each corresponding to a prototype that is discriminative between categories. we retain the corresponding $\boldsymbol{d}$ = 300 motion prototypes from an original set of $\boldsymbol{d'}$ = 12,000 prototypes.

Normalizing position- and- velocity- based features The precise position and size of a cage varies across videos because of variations in the camera angle and the distance between the camera and the cage. To make position- and velocity-based features robust to these variations, these features were normalized with respect to the coordinates of the top, bottom, left and right corners of the cage. These coordinates are now manually annotated, once for each video.

Learning the temporal model SVMHMM combines the advantage of SVM and HMM by discriminatively training models that are similar to hidden Markov models. Here we use a first-order transition model. Given an input sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ of T feature

vectors, the model predicts a sequence of labels $\boldsymbol{y} = (y_1, \ldots, y_T)$ according to the following linear discriminant function:

$$\boldsymbol{y} = argmax_{\boldsymbol{y}} \sum_{t=1}^{T} [\boldsymbol{x}_t \cdot \boldsymbol{w}_{y_t} + \boldsymbol{w}_{t_r}(y_{t-1}, y_t)]$$

$\mathbf{x}_t$ is the concatenation of motion and position-velocity-based feature for frame t of a video sequence, and y_t is the label (one behavior of interest) for frame t. $\mathbf{w}_{y_t}$ is an emission weight vector for the label y_t. $\mathbf{w}_{t_r}$ is a transition weight vector for the transition between the label y_{t-1} and y_t. These weight vectors are learned from 11 training videos in each leave-one-video-out trial. Each training video is split into non-overlapping 1 minute segments, each as a training example $\boldsymbol{X}$.

Extraction of the sub-windows

Motion-features are computed within a sub-window surrounding the animal in order to speed-up the computation and increase accuracy by reducing the chance of matching a space-time motion template with a patch of moving background.

The extraction is described as follows. In the S1/C1 stage, the low-level motion features are computed for the whole frame. The location of the mouse in the C1 map can be accurately computed from the foreground mask (Fig. 2a) knowing the filter size in the S1 stage and the region size of the max-pooling neighborhood in the C1 stage. We then, for each C1 map, crop a small map that is larger than the mouse body with a margin of 5 pixels on each of the four sides (5 pixels on the C1 map is about 20 pixels in the original frame). The template matching in the S2 stage is based on these cropped C1 maps.

Extraction of the bounding boxes

Position- and velocity-based features are derived from the instantaneous location of the animal in a cage. Several features are derived from a bounding box that tightly surrounds the animal in the foreground mask. For example, the width of a mouse is approximated by the width of the bounding box.

Comparison with the system using only motion features. Sometimes different mice behaviors can exhibit very similar motion properties. Contextual information provided by the location of the animal with the cage (e.g., ‘near the food dispenser’) thus becomes

important for disambiguating such behaviors. For example, 'drinking', 'eating', and 'rearing' all have upward motion, but usually occur at different locations. 'Drinking' occurs near the water bottle spout when an animal attaches its mouth to the tip of a drinking tube; 'eating' occurs when an animal reaches the food hopper; and 'rearing' occurs when an animal reaches against the wall. Our solution for removing these ambiguities is to compute a set of 10 position- and velocity-based features such as the distance from a mouse to a drinking tube or food Hooper (see Table S1). To estimate the gain in performance obtained from these 10 features, we evaluated the performance on the *'set B'* using a system that trained only with motion features. Fig. S3 shows the confusion matrix to measure the agreement between the system and human manual scoring. Fig. 3 ('motion + pos') and Fig. S3 ('motion-only') suggest that the addition of the position features to the system benefits 'static' actions most. For instance, the addition of position features improves the accuracy of the system for 'drinking' by 72% and for 'resting' by 29%. The accuracy for the 'eating' behavior also increases by 15%.

Comparison with a benchmark computer vision system. The computer vision system used here for benchmark is the system developed by Dollar, Rabaud, Cottrell, & Belongie at the University of California (San Diego) as part of the *SmartVivarium* project [50]. The system has been shown to outperform several other computer vision systems on several standard computer vision databases and was tested for both the recognition of human and mice behaviors [23]. The authors graciously provided the source code for their system. Training and testing of this system was done in the same way as for our system using a leave-one-video-out procedure on the *'clipped database'*. Here we attempted to maximize the performance of the benchmark system by tuning some of the key parameters such as the number of features and the resolution of the videos used. Nevertheless we found that the default parameters (50 features, a 320x240 video resolution as used for our system) led to the best performance (81% for the system by Dollar et al vs. 93% for our system). It is possible however that further refinement of the corresponding algorithm could nevertheless improve its performance.

SUPPLEMENTARY REFERENCES

31. Mutch, J., Knoblich, U. & Poggio, T. (MIT-CSAIL, 2010).
32. Zorrilla, E.P. et al. Measuring meals: structure of prandial food and water intake of rats. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **288**, 1450-67 (2005).
33. Megens, A.A.H.P., Voeten, J., Rombouts, J., Meert, T.F. & Niemegeers, C.J.E. Behavioural activity of rats measured by a new method based on the piezo-electric principle. *Psychopharmacology* **93**, 382-388 (1987).
34. Casadesus, G., Shukitt-hale, B. & Joseph, J.A. Automated measurement of age-related changes in the locomotor response to environmental novelty and home-cage activity. *Mech. Ageing Dev.* **122**, 1887-1897 (2001).
35. Tamborini, P., Sigg, H. & Zbinden G. Quantitative analysis of rat activity in the home cage by infrared monitoring. Application to the acute toxicity testing of acetanilide and phenylmercuric acetate. *Arch. Toxicol.* **63**, 85-96 (1989).
36. Tang, X. & Sanford, L.D. Home cage activity and activity-based measures of anxiety in 129P3/J, 129X1/SvJ and C57BL/6J mice. *Physiol. Behav.* **84**, 105-15 (2005).
37. Lewejohann, L. et al. Behavioral phenotyping of a murine model of Alzheimer's disease in a seminaturalistic environment using RFID tracking. *Behav. Res. Methods* **41**, 850-856 (2009).
38. Gannon, K.S., Smith, J.C., Henderson, R. & Hendrick, P. A system for studying the microstructure of ingestive behavior in mice. *Physiol. Behav.* **51**, 515-21 (1992).
39. Veeraraghavan, A., Chellappa, R. & Srinivasan, M. Shape-and-behavior encoded tracking of bee dances. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 463-476 (2008).
40. Branson, K. & Belongie, S. Tracking multiple mouse contours (without too many samples). *Proc. IEEE Comp. Vision and Patt. Recogn.* (2005).
41. Spink, A.J., Tegelenbosch, R.A.J., Buma, M.O.S. & Noldus, L.P.J.J. The EthoVision video tracking system-A tool for behavioral phenotyping of transgenic mice. *Physiol. Behav.* **73**, 719-730 (2001).
42. Van Lochem, P.B.A., Buma, M.O.S., Rousseau, J.B.I. & Noldus, L.P.J.J. Automatic recognition of behavioral patterns of rats using video imaging and statistical classification. *Proc. Int. Conf. Measuring Behav.* (1998).
43. Twining, C.J., Taylor, C.J. & Courtney, P. Robust tracking and posture description for laboratory rodents using active shape models. *Behav. Res. Meth. Ins. C.* **33**, 381-391 (2001).
44. Leroy, T., Stroobants, S., Aerts, J.-M., D'Hooge, R. & Berckmans, D. Automatic analysis of altered gait in arylsulphatase A-deficient mice in the open field. *Behav. Res. Methods* **41**, 787-794 (2009).
45. Zurn, J.B., Jiang, X. & Motai, Y. Video-Based Tracking and Incremental Learning Applied to Rodent Behavioral Activity Under Near-Infrared Illumination. *IEEE Trans. Instrum. Meas.* **56**, 2804-2813 (2007).
46. Millecamps, M. et al. Circadian pattern of spontaneous behavior in monarthritic rats: a novel global approach to evaluation of chronic pain and treatment effectiveness. *Arthritis Rheum.* **52**, 3470-8 (2005).
47. De Visser, L., Van Den Bos, R., Kuurman, W.W., Kas, M.J.H. & Spruijt, B.M. Novel approach to the behavioural characterization of inbred mice: automated home cage observations. *Genes Brain Behav.* **5**, 458-66 (2006).

48. Bonasera, S.J., Schenk, A.K., Luxenberg, E.J. & Tecott, L.H. A Novel Method for Automatic Quantification of Psychostimulant-evoked Route-tracing Stereotypy: A pplications to Mus Musculus. *J. Psychopharmacology* **196**, 591-602 (2008).
49. Donohue, K.D., Medonza, D.C., Crane, E.R. & O'Hara, B.F. Assessment of a non-invasive high-throughput classifier for behaviours associated with sleep and wake in mice. *Biomed. Eng. Online* **7**, 14 (2008).
50. Belongie, S., Branson, K., Doll\'ar, P. & Rabaud, V. Monitoring Animal Behavior in the Smart Vivarium. *Proc. Int. Conf. Measuring Behav.* (2005).
51. Xue, X. & Henderson, T.C. Video Based Animal Behavior Analysis From Multiple Cameras. *Proc. IEEE Conf. Multisens. Fusion Integr. Intell. Syst.*, 335-340 (2006).
52. Henderson, T.C. & Xue, X. Constructing Comprehensive Behaviors : A Simulation Study. *Proc. IEEE Conf. Comp. Appl. Ind. Eng.* (2005).
53. Tsochantaridis, I., Hofmann, T., Joachims, T. & Altun, Y. Support Vector Machine Learning for Interdependent and Structured Output Spaces. *Proc. Int. Conf. on Mach. Learn.* (2004).
54. Tsochantaridis, I., Joachims, T., Hofmann, T. & Altun, Y. Large Margin Methods for Structured and Interdependent Output Variables. *J. Mach. Learn. Res.* **6**, 1453-1484 (2005).
55. Vapnik, V. The Nature of Statistical Learning Theory (1995).
56. Weston, J., Elisseeff, A., Scholkopf, B. & Tipping, M. Use of the Zero-Norm with Linear Models and Kernel Methods. *J. Mach. Learn. Res.* **3**, 1439-1461 (2003).